ISAT Reliability

The Question: Are the ISAT assessment tests reliable?

The Importance of the Matter: Statewide assessment tests must be reliable, producing information that is dependable. NCLB language requires documentation of test reliability, as do professional standards in the measurement field. If test scores are unreliable, valid interpretations about student achievements are unattainable.

Methodology of Research: The ISAT contractor, NWEA, routinely collects information about test reliability. This information was reviewed but, in addition, HumRRO conducted its own calculations using a variety of methods. Reliability coefficients were prepared as well as calculations of the standard error of measurement, classification accuracy, and classification consistency. The latter two were needed because the test results are used to classify students into proficiency levels. Such classifications should be precise and replicable.

Calculations were made on the core tests, the blended tests, and on the subscore-reporting units called "Reporting Goals."

Synopsis of Findings: Reliability is determined on a scale from 0 to 1. A measure is considered more reliable the closer it gets to 1. The HumRRO results are similar to those obtained by NWEA. The overall test reliabilities were all above 0.80 and most were around 0.86. The total test standard error of measurement was about 3 points. For the subtests (e.g., Number Sense or Literal Comprehension) where there are fewer test items, the reliability decreases substantially to about 0.50 and the standard errors of measurement increase to about 6-7 points. In other words, the total test score is the most stable and the subtest scores are less stable. These values are typical for statewide assessment tests.

Standard errors of measurement also were calculated using Rasch Item Response Theory methods wherein a standard error of measurement is calculated for each test item rather than for the overall test. The NWEA and HumRRO results are almost identical.

Accuracy addresses the matter of whether the test score accurately classifies a student, just as one might look at a thermometer and consider whether or not the indicated temperature is "accurate." The accuracy indices were in the 0.75 – 0.85 range, with the blended test yielding higher values.

Consistency describes the likelihood that the student would have attained the same proficiency classification on a second administration of a parallel form of the test. The results showed that the consistency indices were in the range of 0.65 – 0.79 with the blended test yielding higher values.

For both accuracy and consistency, if the analyses consider only the Proficiency/Not Proficient dichotomy, the accuracy and consistency values increase to the 0.88 –0.95 range.

Implications for Future Direction: The values obtained in these analyses are typical. The easiest way to increase reliability is to increase the number of items on the tests and subtests. For example, increasing subtests from 6 items to 10 items would increase the reliability of the subscore. Similarly, increasing the overall length of the tests would increase overall reliability.

Reliability should be monitored with each administration of the ISAT and adjustments made in the test structure when necessary.

In its score interpretation guides, the State should reinforce the psychometric principle that the most reliable scores at the individual student level are those derived from the total test, and the subgoal reports are only “advisory.” On the other hand, when aggregating data across students into school and district units, the subgoal reports can be considered to be more reliable.

Report: Idaho Standards Achievement Test: Independent Calculations of Reliability Estimates, Standard Errors of Measurement, Classification Accuracy, and Classification Consistency